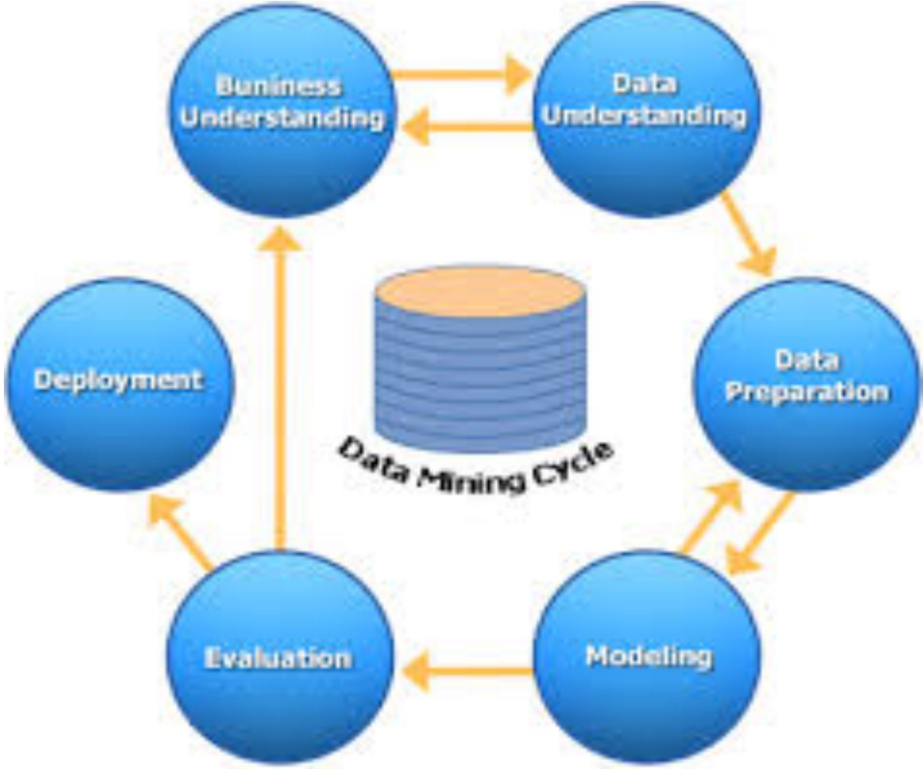
Describing Data

Data mining process

CRISP-DM (Cross Industry Standard Process for Data Mining) provides a process model for data mining that consists of six major phases

- Business understanding
- Data understanding
- Data preparation
- Modeling
- Evaluation
- Deployment



http://www.crisp-dm.org/Process/index.htm

Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Assess Situation
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

Select Data
Rationale for Inclusion/ Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data

Format Data
Reformatted Data

Dataset
Dataset Description

Modeling

Select Modeling Techniques
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Descriptions

Assess Model
Model Assessment
Revised Parameter Settings

Evaluation

Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

Deployment

Plan Deployment
Deployment Plan

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

Figure 3: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model

What is data?

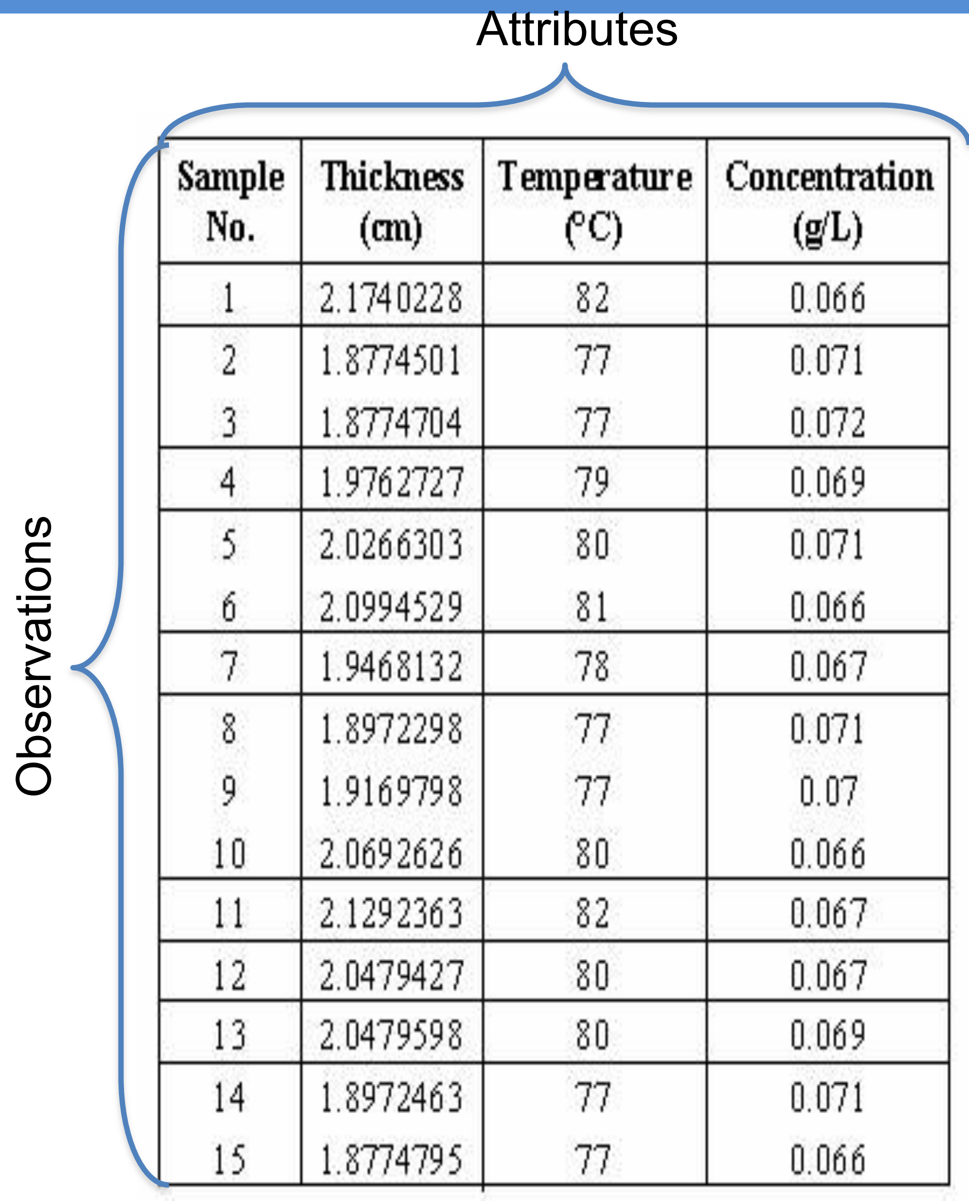
- In today's world centralized around information technology, huge amounts of data are produced and stored each day. Often these data come from automatic detection system, sensors, and scientific instrumentation, or you produce them daily and unconsciously every time you make a transaction on web, you create a data.

- But what is data? The data actually are not information, at least in terms of the forms. In the formless stream of bytes, at first glance it is difficult to understand their essence if not strictly the number, word, or time that they report.

- Information is actually the result of preprocessing, which taking into account a certain set of data, extract some conclusions that can be used in various ways. This process of extracting information from the raw data is precisely data analysis.

What is data?

- Collection of data objects and their attributes/variable, characteristic or feature
- An attribute is a property of an object, e.g., height of a person, or temperature of furnace
- A collection of features describe an object. Object is also known as record, observation, sample etc.
- Data is arranged in a table. It simply arranges data in a convenient form



Attributes

Observations

Sample No.	Thickness (cm)	Temperature (°C)	Concentration (g/L)
1	2.1740228	82	0.066
2	1.8774501	77	0.071
3	1.8774704	77	0.072
4	1.9762727	79	0.069
5	2.0266303	80	0.071
6	2.0994529	81	0.066
7	1.9468132	78	0.067
8	1.8972298	77	0.071
9	1.9169798	77	0.07
10	2.0692626	80	0.066
11	2.1292363	82	0.067
12	2.0479427	80	0.067
13	2.0479598	80	0.069
14	1.8972463	77	0.071
15	1.8774795	77	0.066

Data Representation

- Data has form: $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ (labeled), or $\{x_1,\ldots,x_n\}$ (unlabeled)
- What the label y looks like is task-specific
- What about x which denotes a real-world object (e.g., image or text document)?
- Each example x is a set of (numeric) features/attributes/dimensions
- Features encode properties of the object which x represents
- x is commonly represented as a D × 1 vector
- Representing a 28 × 28 image: x can be a 784 × 1 vector of pixel values
- Representing a text document: x can be a vector of word-counts of words appearing in that document

Types of Data

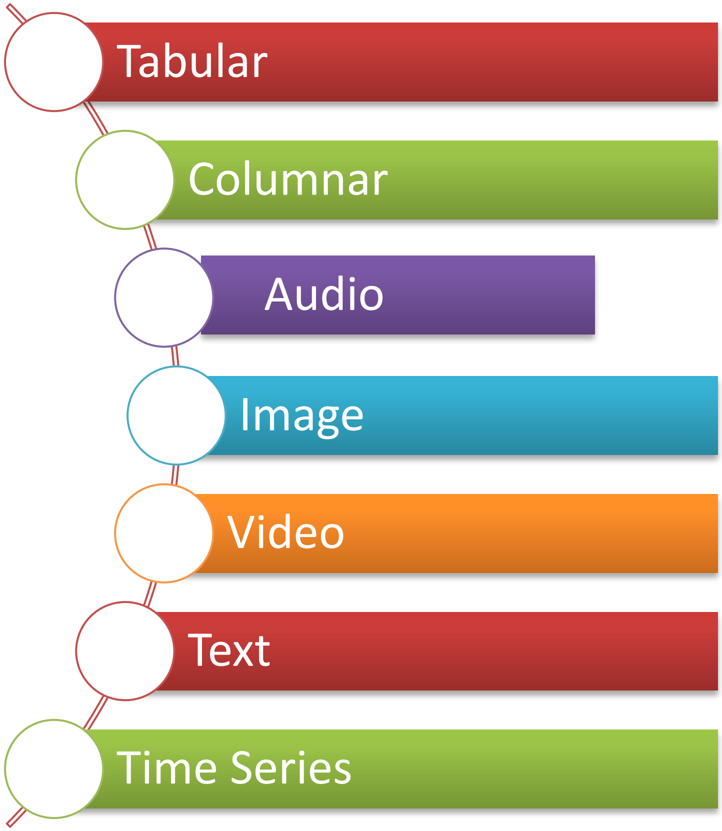
- **Qualitative/Categorical variable:** When a variable can be placed into well-defined groups or categories, that does not depend on order. It can take only specific value, e.g., color of a car or gender. There are two types of categorical values: nominal and ordinal
- **Nominal**: describe a variable with limited number of different values that can not be ordered. (ID number, color, industry type (financial, engineering, retail)), Gender (Male, Female)
- **Ordinal**: a variable whose value can be ordered or ranked (height:{tall, med, short})

Types of Data

- **Numerical/Quantitative variable:** are values or observations that come from measurements. There are two types of different values: discrete and continuous numbers. Discrete values are values that can be counted and that are distinct and separated from each other. Continuous values, on the other hand, are values produced by measurements or observations that assume any value within a defined range.

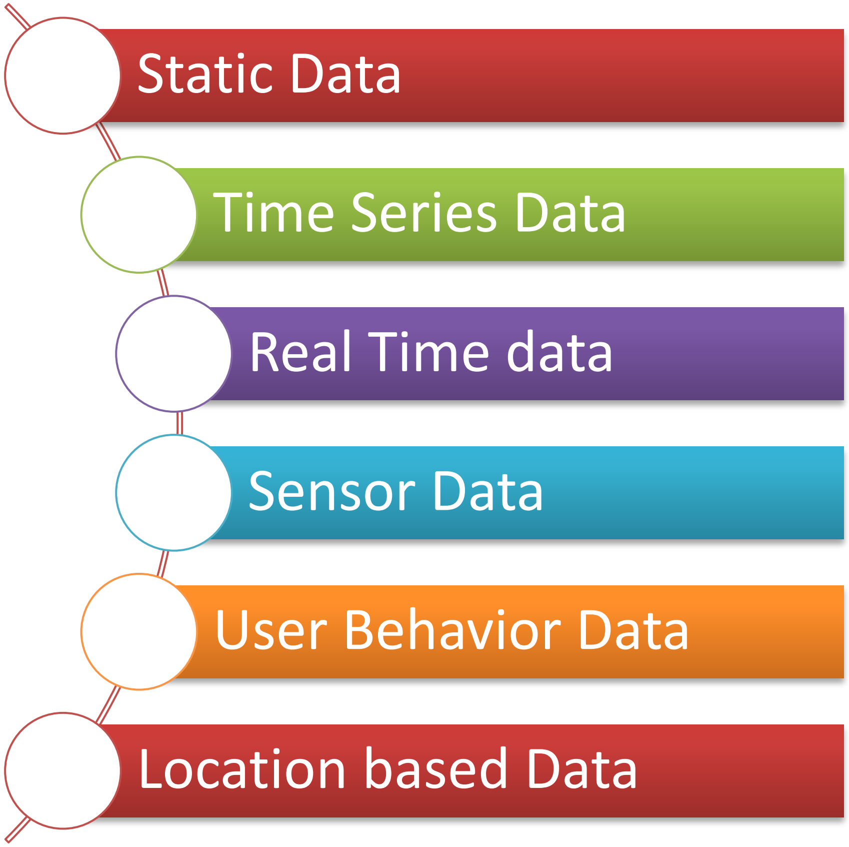
Types of data sets

Data Format



- Tabular
- Columnar
- Audio
- Image
- Video
- Text
- Time Series

What does Data Represent

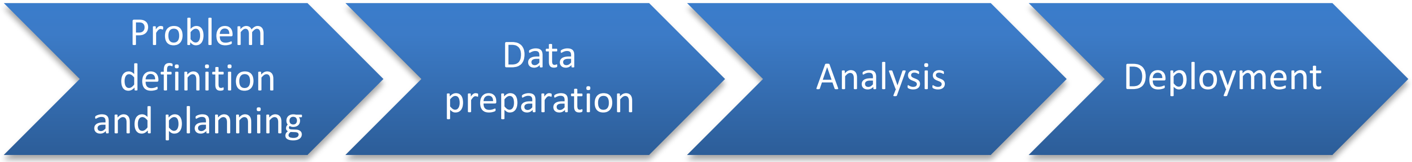


- Static Data
- Time Series Data
- Real Time data
- Sensor Data
- User Behavior Data
- Location based Data

You got your data: what’s next:

What kind of analysis do you need which model is more appropriate for it? …

Data Analysis



- The starting point for data analysis is a data set which contains the measured or collected data values represented as numbers or text. The data is raw before it has been transformed or modified.

- All the disciplines collect data about items that are important to that field. These items are organized into a table for data analysis where each row , referred to as an observation, contains information about the specific item. Data table also contain information about the object and is known as attribute.

Understanding the Nature of the Data

•The object of study of the data analysis is basically the data. The data then will be the key players in all processes of the data analysis. They constitute the raw material to be processed, and thanks to their processing and analysis it is possible to extract a variety of information in order to increase the level of knowledge of the system under study, that is, one from which the data came from.

When the Data Become Information

•Data are the events recorded in the world. Anything that can be measured or even categorized can be converted into data. Once collected, these data can be studied and analyzed both to understand the nature of the events and very often also to make predictions or at least to make informed decisions.

When the Information Becomes Knowledge

- You can speak of knowledge when the information is converted into a set of rules that help you to better understand certain mechanisms and so consequently, to make predictions on the evolution of some events.

Data Analysis

- The purpose of data analysis is precisely to extract information that is not easily deducible but that, when understood, leads to the possibility of carrying out studies on the mechanisms of the systems that have produced them, thus allowing the possibility of making forecasts of possible responses of the systems and their solution in time.
- Starting from a simple methodical approach on data protection, data analysis has become a real discipline leading to the development of real methodologies generating models.
- The model is in fact the translation into mathematical form of a system. Once there is a mathematical form able to describe system responses under different levels of precision, you can then make predictions about its responses to certain unseen/new inputs. Thus, the aim of data analysis is not only the model, but the goodness of its predictive power.

Data Analysis

- The predictive power of a model depends not only on the quality of the modeling technique but also on the ability to choose a good dataset upon which to build the model.
- So the search for the data, their extraction, and their subsequent preparation belong to the data analysis because of their importance in the success of the results.
- In parallel to all stages of processing of data analysis, various methods of data visualization have been developed.
- In fact to understand the data, both individually and in terms of the role they play in the entire data set, there is no better system than to develop the techniques of graphic representation capable of transforming information in figures which help more easily to understand their meaning.

Data Analysis

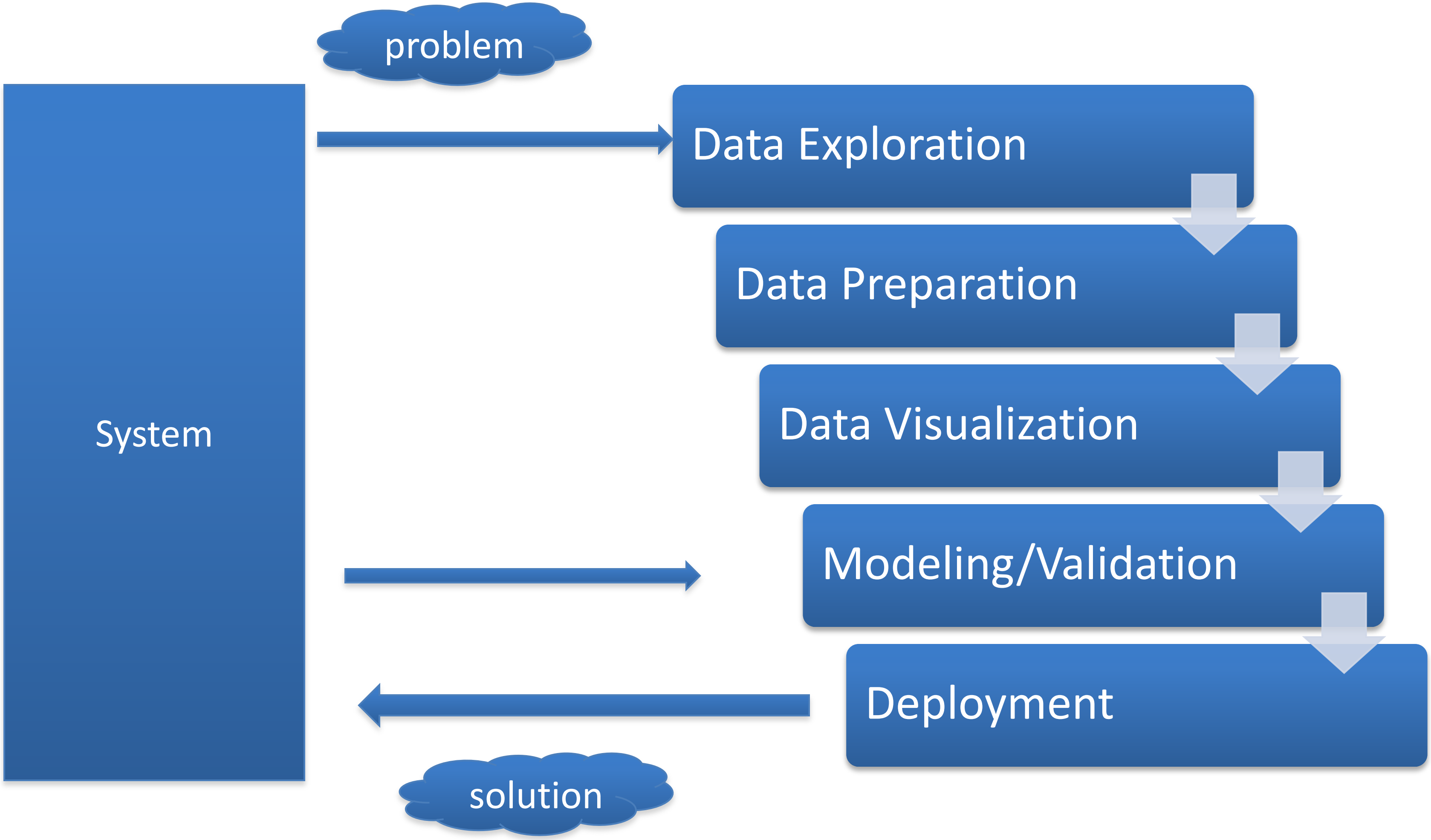
- Over the years lots of display modes have been developed for different modes of data display: the charts
- At the end of the data analysis, you will have a model and a set of graphical displays and then you will be able to predict the responses of the system under study; after that you will move to the test phase. The model will be tested using new data/unseen data.
- Once the model has been assessed, you can move to the last phase of data analysis – deployment.
- Data analysis is a discipline that is well suited to many professional activities. So, knowledge of what it is and how it can be put into practice will be relevant for consolidating the decisions to be made. It will allow us to test hypotheses, and to understand more deeply the systems analyzed.

Data Analysis

Data analysis can be described as a process consisting of several steps in which the raw data are transformed and processed in order to produce data visualizations and can make predictions thanks to a mathematical model based on the collected data. Then, data analysis is nothing more than a sequence of steps, each of which plays a key role in the subsequent ones. So, data analysis is almost schematized as a process chain consisting of the following sequence of stages:

- Problem definition
- Data extraction
- Data cleaning
- Data transformation
- Data exploration
- Predictive modeling
- Model validation/test
- Visualization and interpretation of results
- Deployment of the solution

Data Analysis Process



problem

System

Data Exploration

Data Preparation

Data Visualization

Modeling/Validation

Deployment

solution

Data Analysis

Problem Definition. The process of data analysis actually begins long before the collection of raw data. In fact, a data analysis always starts with a problem to be solved, which needs to be defined.

The problem is defined only after you have well-focused the system you want to study: this may be a mechanism, an application, or a process in general. Generally this study can be in order to better understand its operation, but in particular the study will be designed to understand the principles of its behavior in order to be able to make predictions, or to make choices (defined as an informed choice).

Data Analysis

The definition step and the corresponding documentation (*deliverables*) of the scientific problem or business are both very important in order to focus the entire analysis strictly on getting results. In fact, a comprehensive or exhaustive study of the system is sometimes complex and you do not always have enough information to start with. So the definition of the problem and especially its planning can determine uniquely the guidelines to follow for the whole project.

Once the problem has been defined and documented, you can move to the **project planning** of a data analysis. Planning is needed to understand which professionals and resources are necessary to meet the requirements to carry out the project as efficiently as possible. So you're going to consider the issues in the area involving the resolution of the problem. You will look for specialists in various areas of interest and finally install the software needed to perform the data analysis.

Data Analysis

Thus, during the planning phase, the choice of an effective team takes place. Generally, these teams should be cross-disciplinary in order to solve the problem by looking at the data from different perspectives. So, the choice of a good team is certainly one of the key factors leading to success in data analysis.

Data Analysis

Data Extraction. Once the problem has been defined, the next step is to obtain the data in order to perform the analysis. The data must be chosen with the basic purpose of building the predictive model, and so their selection is crucial for the success of the analysis as well. The sample data collected must reflect as much as possible the real world, that is, how the system responds to stimuli from the real world. In fact, even using huge data sets of raw data, often, if they are not collected competently, these may portray false or unbalanced situations compared to the actual ones.

Thus, a poor choice of data, or even performing analysis on a data set which is not perfectly representative of the system, will lead to models that will move away from the system under study.

Data Analysis

The search and retrieval of data often require a form of intuition that goes beyond the mere technical research and data extraction. It also requires a careful understanding of the nature of the data and their form, which only good experience and knowledge in the problem's application field can give.

Regardless of the quality and quantity of data needed, another issue is the search and the correct choice of **data sources**.

When you want to get the data, a good place to start is just the Web. But most of the data on the Web can be difficult to capture; in fact, not all data are available in a file or database, but can be more or less implicitly content that is inside HTML pages in many different formats.

Data Analysis

To this end, a methodology called **Web Scraping**, which allows the collection of data through the recognition of specific occurrence of HTML tags within the web pages, has been developed. There are software specifically designed for this purpose, and once an occurrence is found, they extract the desired data. Once the search is complete, you will get a list of data ready to be subjected to the data analysis.

Data Analysis

Data Preparation. Among all the steps involved in data analysis, data preparation, though seemingly less problematic, is in fact one that requires more resources and more time to be completed. The collected data are often collected from different data sources, each of which will have the data in it with a different representation and format. So, all of these data will have to be prepared for the process of data analysis.

The preparation of the data is concerned with obtaining, cleaning, normalizing, and transforming data into an optimized data set, that is, in a prepared format, normally tabular, suitable for the methods of analysis that have been scheduled during the design phase.

Many problems that must be avoided, such as invalid, ambiguous, or missing values, replicated fields, or out-of-range data.

Data Analysis

Data Exploration/Visualization

Exploring the data is essentially the search for data in a graphical or statistical presentation in order to find patterns, connections, and relationships in the data. Data visualization is the best tool to highlight possible patterns.

In recent years, data visualization has been developed to such an extent that it has become a real discipline in itself. In fact, numerous technologies are utilized exclusively for the display of data, and equally many are the types of display applied to extract the best possible information from a data set.

Data Analysis

Data exploration consists of a preliminary examination of the data, which is important for understanding the type of information that has been collected and what they mean. In combination with the information acquired during the definition problem, this categorization will determine which method of data analysis will be most suitable for arriving at a model definition.

Generally, this phase, in addition to a detailed study of charts through the visualization data, may consist of one or more of the following activities:

- Summarizing data
- Grouping data
- Exploration of the relationship between the various attributes
- Identification of patterns and trends Construction of regression/classification models

Data Analysis

Generally, the data analysis requires processes of summarization of statements regarding the data to be studied. The **summarization** is a process by which data are reduced to interpretation without sacrificing important information.

Clustering is a method of data analysis that is used to find groups united by common attributes (**grouping**).

Another important step of the analysis focuses on the **identification** of relationships, trends, and anomalies in the data. In order to find out this kind of information, one often has to resort to the tools as well as performing another round of data analysis, this time on the data visualization itself.

Other methods of data mining, such as decision trees and association rules, automatically extract important facts or rules from data. These approaches can be used in parallel with the data visualization to find information about the relationships between the data.

Data Analysis

Predictive Modeling

Predictive modeling is a process used in data analysis to create or choose a suitable statistical model to predict the probability of a result.

After exploring data you have all the information needed to develop the mathematical model that encodes the relationship between the data. These models are useful for understanding the system under study, and in a specific way they are used for two main purposes. The first is to make predictions about the data values produced by the system; in this case, you will be dealing with **regression models**. The second is to classify new data products, and in this case, you will be using **classification models** or **clustering models**.

Data Analysis

In fact, it is possible to divide the models according to the type of result that they produce:

Classification models: If the result obtained by the model type is categorical.

Regression models: If the result obtained by the model type is numeric.

Clustering models: If the result obtained by the model type is descriptive.

Model Validation

Validation of the model, that is, the test phase, is an important phase that allows you to validate the model built on the basis of starting data. That is important because it allows you to assess the validity of the data produced by the model by comparing them directly with the actual system. But this time, you are coming out from the set of starting data on which the entire analysis has been established.

Data Analysis

Deployment

This is the final step of the analysis process, which aims to present the results, that is, the conclusions of the analysis. In the deployment process, in the business environment, the analysis is translated into a benefit for the client who has commissioned it. In technical or scientific environments, it is translated into design solutions or scientific publications. That is, the deployment basically consists of putting into practice the results obtained from the data analysis

There are several ways to deploy the results of a data analysis or data mining. Normally, a data analyst's deployment consists in writing a report for management or for the customer who requested the analysis. This document will conceptually describe the results obtained from the analysis of data. The report should be directed to the managers, who are then able to make decisions. Then, they will really put into practice the conclusions of the analysis.

Data Analysis

In the documentation supplied by the analyst, each of these four topics will generally be discussed in detail:

- Analysis results
- Decision deployment
- Risk analysis
- Measuring the business impact

When the results of the project include the generation of predictive models, these models can be deployed as a stand-alone application or can be integrated within other software.

- Preparing data also prepares the scientist so that when using prepared data the scientist produces better models, and faster.

- Good data is a prerequisite for producing effective models of any type.

- Several data mining methods are sensitive to the scale and/or
type of the variables

- Different variables (columns of our data sets) may have rather
different scales

- Some methods are not able to handle either nominal or numeric
Variables

- We may need to “create” new variables to achieve our objectives

- Sometimes we are more interested in relative values (variations) than absolute values

- We may be aware of some domain-specific mathematical relationship among two or more variables that is important for the task

- Frequently we have data sets with unknown variable values

- Our data set may be too large for some methods to be applicable

Cont’d

- These techniques are not mutually exclusive; they may work together. For example, data cleaning can involve transformations to correct wrong data, such as by transforming all entries for a date field to a common format.

- Data cleaning techniques, when applied before mining, can substantially improve the overall quality of the modeling.

- Data cleaning routines work to “clean” the data by filling in missing values, smoothing noisy data, identifying o removing outliers, and resolving inconsistencies.

- If the user believe that the data are dirty, they are unlikely to trust the results of any data mining that has been applied to it. Furthermore, dirty data can cause confusion for the mining procedure, resulting in unreliable output.

Measure of Data Quality

- Raw data is often not useful without some kind of organization or manipulation. Raw data seems to be just a bunch of meaningless values without any context or some level of organization.
- In recent years, data quality has gained more an more attention due to extended use of data warehouse systems and a higher relevance of customer relationship management.
- Due to this fact for decision makers, the benefits of data depends heavily on their completeness, correctness, and timeliness, respectively. Such properties are known as data quality dimensions.
- The consequences of poor data quality are manifold: They range from worsening customer relationships and customer satisfaction by falsely addressing customers to insufficient decision support for managers.

Measure of Data Quality

- Data quality dimension has been used to describe the measure of the quality of data. However, even amongst data quality professionals the key data quality dimensions are not universally agreed.
- A data quality dimension is different to, and should not be confused with other dimension terminologies such as those used in:
  - Other aspects of data management e.g., a data warehouse dimension or a data cube dimension

Measure of Data Quality

- There are six best practice definitions as generic data quality dimensions. These dimensions should be adopted by data quality practitioners as the standard method for assessing and describing the quality of data. However, in some cases or situations one or more dimension may not be relevant.

- Before we use these dimensions, one needs to agree the quality rules against which the data needs to be assessed against. These rules should be developed based upon the data quality dimensions discussed later, organizational requirements for data and the impact on an organization of data not complying with these rules.

Measure of Data Quality

- Organizations select the data quality dimensions and associated dimension threshold based on their business context, requirements, level of risks etc.

Measure of Data Quality

- The following measure can be used to test the quality of data
  - *Completeness: The proportion of stored data against the potential of “100% complete”*
  - *Uniqueness: No observation will be recorded more than once based upon how that observation is identified.*
  - *Velocity: The rate at which data is coming especially for streaming data*
  - *Accuracy: The degree to which data correctly describes the “real world” event being described*
  - *Consistency: The absence of difference, when comparing two or more representations of a thing against a definition.*
  - *Accessibility: How easy it is to access the data.*

Data Quality

- What kind of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?
- Examples of data quality problems:

  Missing Values
  Noise and Outliers
  Duplicate values
  Inconsistent Dates
  Impossible values (Negative Sales)
  check using if condition

Data Analysis

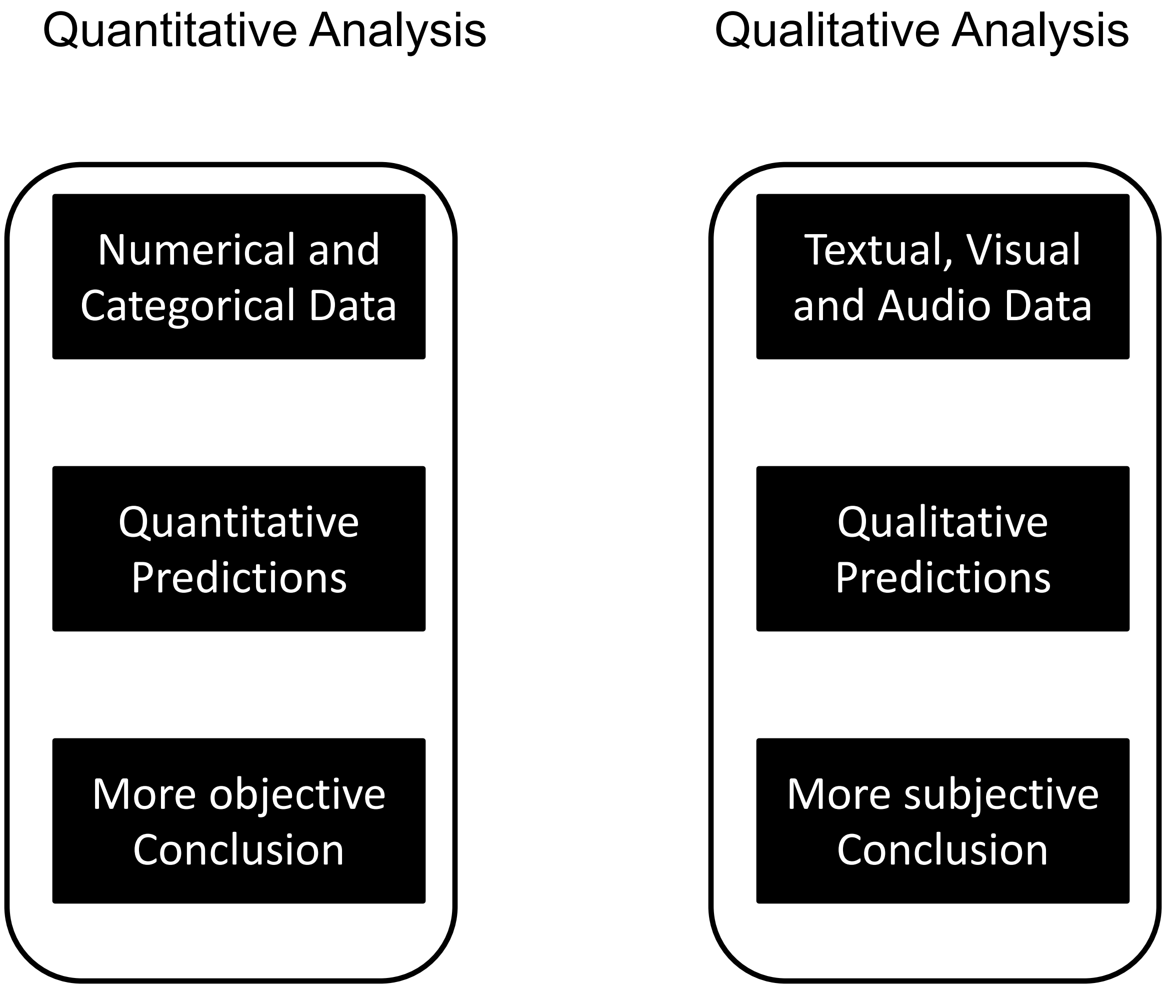
Quantitative and Qualitative Data Analysis

Data analysis is therefore a process completely focused on data, and, depending on the nature of the data, it is possible to make some distinctions.

When the analyzed data have a strictly numerical or categorical structure, then you are talking about **quantitative analysis**, but when you are dealing with values that are expressed through descriptions in natural language, then you are talking about **qualitative analysis**.

Precisely because of the different nature of the data processed by the two types of analyses, you can observe some differences between them.

TYPES OF ANALYSIS



Quantitative Analysis

Numerical and Categorical Data

Quantitative Predictions

More objective Conclusion

Qualitative Analysis

Textual, Visual and Audio Data

Qualitative Predictions

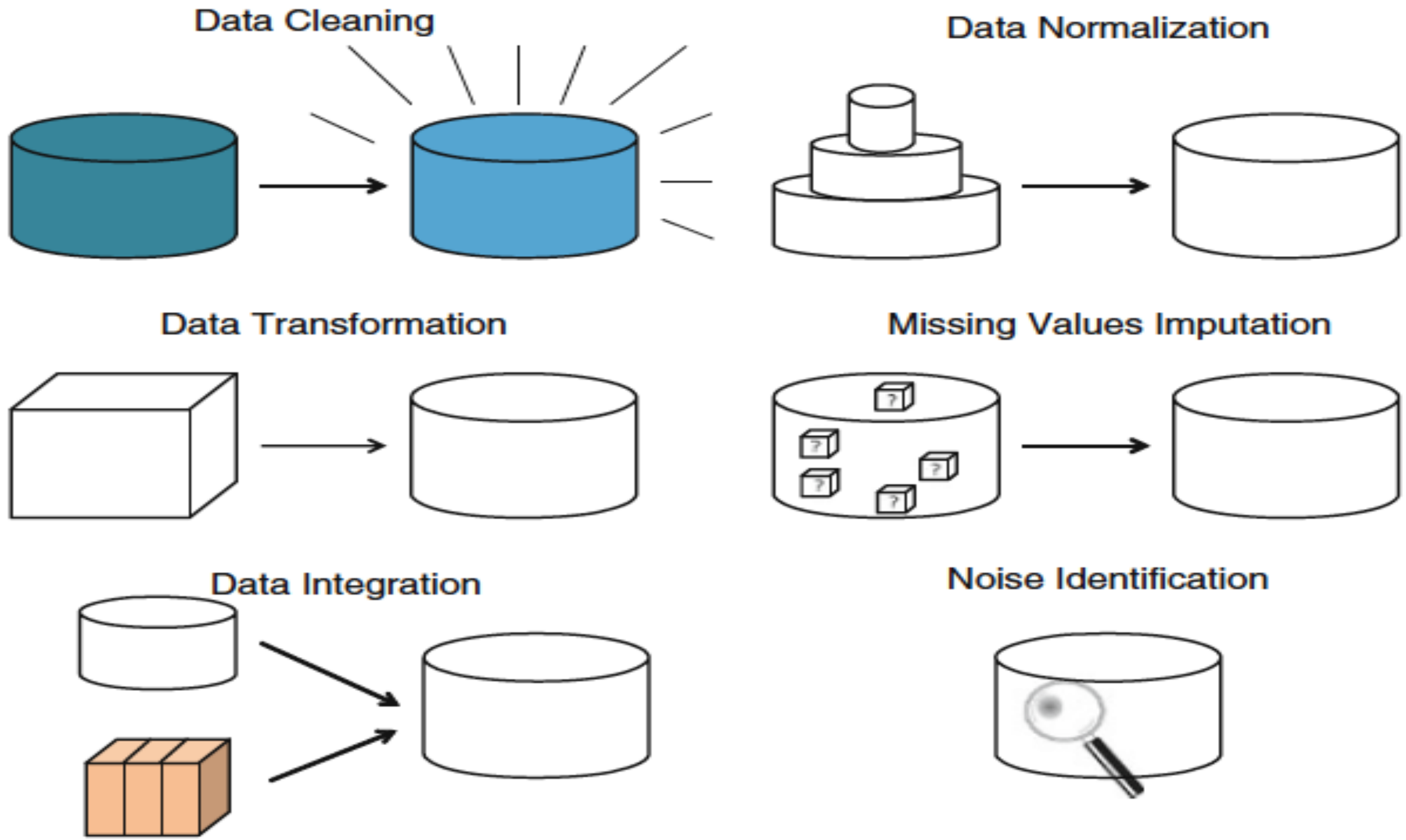
More subjective Conclusion

Major tasks involved

- **Integration of data**
  - Integration of data from multiple databases, or files
- **Data discretization (for numerical data)**
- **Data Cleaning**
  - Fill in missing values, smooth noisy data, remove outliers and resolve inconsistencies
- **Data Transformation**
  - Scaling/normalization and aggregation. Normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements.
- **Data reduction**
  - Optimize the features/attributes and obtain reduced representation in volume.

Data Analysis

- Forms of data preparation



Data Sets

The Linked Open Data Cloud

To give an idea of open data sources available online, you can look at the LOD cloud diagram (http://lod-cloud.net), which displays all the connections of the data link between several open data sources

currently available in the network